DOCUMENT RESUME

ED 213 730 TM 820 074

AUTHOR Wilcox, Rand R.

TITLE Achievement Tests and Latent Structure Models. Studies in Measurement and Methodology, Work Unit 3: Psychometric Problems in Achievement Tests.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 79
GRANT OB-NIE-G-78-0213
NOTE 27p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Guessing (Tests); *Mathematical Models; Multiple Choice Tests; *Scoring Formulas; Testing; Test Theory
IDENTIFIERS *Latent Structure Analysis

ABSTRACT

In the past, several latent structure models have been proposed for handling problems associated with measuring the achievement of examinees. Typically, however, these models describe a specific examinee in terms of an item domain or they describe a few items in terms of a population of examinees. In this paper, a model is proposed which allows a population of examinees to be described in terms of an item domain. As is illustrated, the model yields a new formula score which allows guessing to vary over the population of examinees. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED213730

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Deliverable: May 1979

Studies in Measurement and Methodology

ACHIEVEMENT TESTS AND LATENT STRUCTURE MODELS

Rand R. Wilcox

Work Unit 3: Psychometric Problems in Achievement Tests

Rand R. Wilcox, Project Director

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. S. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Grant Number
OB-NIE-G-78-0213

TM 820 074

Center for the Study of Evaluation
Graduate School of Education
University of California - Los Angeles

The work upon which this publication is based was performed pursuant to a contract with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

ABSTRACT

In the past, several latent structure models have been proposed for handling problems associated with measuring the achievement of examinees. Typically, however, these models describe a specific examinee in terms of an item domain or they describe a few items in terms of a population of examinees. In this paper, a model is proposed which allows a population of examinees to be described in terms of an item domain. As is illustrated, the model yields a new formula score which allows guessing to vary over the population of examinees.

Consider a specific instructional program which is intended to promote certain skills in the students taking the course. When measuring the achievement of an examinee, the question arises as to the proportion of skills the examinee has acquired from among all those that are being taught. The simplest approach to this problem is to administer n randomly selected test items and to use the observed score as an indication of what the examinee has learned. Since observed scores contain error, it is natural to search for a procedure that reflects more accurately what the test measures.

Consider, for instance, a multiple-choice test item having t alternatives from which to choose, one of which is designated as being correct. Since an examinee who has not acquired the skill might give a correct response, the observed score is at best a crude indication of whether the skill has been learned.

How do we correct this problem and what advantages do these corrections have? One well-known approach is to use the correction for guessing formula (see, e.g., Lord and Novick, 1968, Chapter 14) which assumes that examinees who do not know guess at random. (For a recent modification see Abu - Sayf, 1977.) However, this approach has been criticized from several points of view. First, examinees might be able to eliminate one or more of the alternatives on a multiple-choice test item without knowing the correct response. Second, some examinees might have misinformation, i.e., they would, if given the chance, always choose the same incorrect response on independent trials because they believe their choice is indeed correct. Still another possible objection is that if there are no omitted responses, the corrected for guessing formula score is perfectly correlated with the number-right score. Lyerly (1951),

for example, criticized the approach taken by Hamilton (1950) for this reason. However, Chernoff (1962) argues that Lyerly's criticism is an oversimplification. He goes on to suggest a scoring procedure based on the assumption of random guessing and the criterion of minimum mean squared error. Nevertheless, while there might be situations where examinees guess approximately at random, it would seem, as pointed out by Lord and Novick (1968, p. 309) that this assumption can seldom be seriously entertained. However, an argument in favor of the formula score can still be made (Lord, 1975). For a review of some of the other issues involved, the reader is referred to Diamond and Evans (1973).

Another possible solution to this problem is the application of latent structure models that are designed specifically for describing achievement test items. Several models have been proposed to handle various situations (e.g., Brownless and Keats, 1958; Marks and Noll, 1967; White and Clark, 1973; Duncan, 1974; Proctor, 1970; Dayton and Macready, 1976; Macready and Dayton, 1977; Wilcox, 1977a, 1977b; Knapp, 1977). One inadequacy of existing models is that they tend to characterize a single skill in terms of a population of examinees or they characterize a domain of skills in terms of a single examinee. The purpose of the present paper is to extend some of these models to a situation where we describe a population of examinees in terms of a domain of skills. As will be seen, the proposed model yields a new formula score which allows guessing to vary over the population of examinees. We begin by describing the general framework within which we propose to work.

We assume that:

1. There is a large (possibly infinite) domain of achievement test items and that we are interested in characterizing a population of examinees in terms of these items.

2. Each item is scored as a correct (1) or incorrect (0) response.

3. It is meaningful to say that an examinee either "knows" or "does not know" the correct response. We do not assume that learning is all or none as is done by Macready and Dayton (1977). This is not to say that the all or none view is incorrect; the point is that this particular assumption about learning is not required in order to say that an examinee either knows or does not know the correct response. (See Wilcox, in press.)

4. Examinees in either of the two latent states might give an inappropriate response. That is, examinees who know might give an incorrect response because they misinterpret the question, get distracted, etc. Also, examinees who do not know might guess the correct response.

For a particular examinee, let ζ denote the proportion of items that he/she knows, let α represent the conditional probability of knowing and being incorrect and let β denote the probability of guessing the correct response given that the examinee does not know. Thus, in terms of a single test item, an examinee belongs to one of four mutually exclusive events as depicted in Table 1. The cell entries represent the corresponding probabilities. The first step in using the model described here is estimating ζ for a particular examinee. One possible solution is described by Wilcox (in press) in which a random sample of items is administered on two occasions not widely separated in time. It should be noted however, that considerable care must be exercised when using this approach. As an extreme case, if we administer only one item to an examinee, wait five minutes, and ask the examinee to respond to the item again, it seems likely that he would

respond in the same fashion even if he does not know the correct response. Certainly steps can be taken to minimize this problem. For example, we might administer two n-item tests on seperate days and repeat only a few of the items on the second day.

One problem with this method of obtaining test-retest information on items is that it may be administratively difficult to obtain results on more than a few items. We describe, therefore, a slight variation of this solution which might be used. Rather than administer the same item on two occasions in time, one might use pairs of equivalent items instead. By equivalent items it is meant that the examinee knows the answer to both items or to neither one. It is also assumed that responses are independent among a particular class of items. This is the assumption of local independence which is typically made when using latent structure models (Lazarsfeld and Henry, 1968, p. 22). This assumption might appear to be restrictive but it is certainly more realistic than assuming examinees guess at random as many psychometricians have done. Anderson (1959, p. 11) argues in favor of this assumption on the grounds that the latent quantities are the only important factors and that once these are determined behavior is random in the sense of statistical independence.

The four probabilities associated with the four possible outcomes for an examinee answering an equivalent pair of items are:

$$p_{11} = (1-\alpha)^2\,\zeta+\beta^2(1-\zeta)$$

$$p_{10} = p_{01} = (1-\alpha)\alpha\zeta+\beta(1-\beta)(1-\zeta)$$

$$p_{00} = \alpha^2\zeta+(1-\beta)^2(1-\zeta),$$

The subscripts on the parameter p indicate a particular pattern of correct and incorrect responses. Thus, p_{10} is the probability of getting a correct on the first item of an equivalent item pair and an incorrect on the other. Note that the assumption $p_{10}=p_{01}$ can be tested with McNemar's test (McNemar, 1947). Wilcox (1977a) gives some results on the power function of this test that might be useful when checking this requirement of the model.

For n randomly sampled item pairs administered to a single examinee, the observed frequencies of correct and incorrect responses can be represented in a two-by-two table:

		Item 2: 1	Item 2: 0	
Item 1	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n

Thus, there are a item pairs for which the examinee got both items correct.

For the two important special cases $\alpha=0$, $\beta>0$ and $\alpha>0$, $\beta=0$, exact solutions for ζ are possible. If, for example, $\alpha=0$,

$$\hat{\zeta} = 1-nd^{-1}[(b+c)/2n+d/n]^2$$

and

$$\hat{\beta} = (b+c)/(b+c+2d)$$

are maximum likelihood estimates of ζ and β. Note that these statistics use $(b+c)/2n$ to estimate $p=p_{10}=p_{01}$ which is unbiased, maximum likelihood, and efficient (Wilcox, 1977a). If $\beta=0$ maximum likelihood estimates are

$$\hat{\zeta} = (4an)^{-1}\ (2a+b+c)^2$$

and

$$\hat{\alpha} = (b+c)(2a+b+c)^{-1}.$$

Results reported by Macready and Dayton (1977) suggest that $\beta \approx 0$ for the examinees in their study while the assumption $\alpha=0$ might be reasonable when multiple choice items are used.

When the assumption $p_{10}=p_{01}$ is contraindicated by the data, the following modification might be used. Suppose $\alpha=0$. Given that the examinee does not know the correct response to a randomly selected pair of equivalent items, let β_1 be the probability of guessing the correct response to the first item and let β_2 be the probability of guessing for the second item in the item pair. In this case

$$p_{11} = \zeta+\beta_1\beta_2(1-\zeta)$$

$$p_{10} = \beta_1(1-\beta_2)(1-\zeta)$$

$$p_{01} = (1-\beta_1)\beta_2(1-\zeta)$$

$$p_{00} = (1-\beta_1)(1-\beta_2)(1-\zeta)$$

It follows that

$$\hat{\beta}_1 = b/(b+d)$$

$$\hat{\beta}_2 = c/(c+d)$$

and

$$\hat{\zeta} = 1-((c+d)/d)((b+d)/n)$$

are maximum likelihood estimates of β_1, β_2 and ζ respectively (Wilcox, 1977b). For $\beta=0$, a similar model can be derived which corresponds to Wilcox's first modification of a model proposed by Goodman and Kruskal (Wilcox, 1977b).

As just indicated existing solutions to certain latent structure models can be modified to estimate ζ in terms of a single examinee and an item domain. The importance of these estimates in the present paper will become clear. We

now turn attention to our main results. Only situations where either $\alpha=0$ or $\beta=0$ are considered. The reason for this is that these special cases yield exact estimates of the parameters in the model to be described. Such cases are important because we might obtain good approximations to more general models. From a statistical point of view, approximations are of interest since they provide initial trial values in iterative estimation procedures when a more general model is preferred. If arbitrary initial values are used, estimation procedures might converge to the wrong value (Goodman, 1974). (See, also, Barnett, 1966; Kale, 1962.) The results reported here do not guarantee that this problem will be eliminated but at a minimum they do have the practical advantage of possibly reducing the number of iterations when an iterative estimation procedure must be used.

The Dirichlet-Multinomial Model

When a particular examinee responds to a randomly chosen item and when $\alpha=0$, the examinee belongs to one of three disjoint sets:

1. The examinee knows the correct response and gives it
2. The examinee does not know and guesses correctly. The probability of this event is $(1-\zeta)\beta$ which for notational convenience is written as γ.

3. The examinee does not know and gives the correct response.

Consider n randomly sampled items. Let y denote the number of items the examinee knows and let z be the number of items for which the examinee guesses the correct response. Then x, the examinee's observed (number correct) score is simply

$$x = y + z.$$

For the assumptions made, y and z have a multinomial distribution given by

$$\frac{n!\ \zeta^{y}\gamma^{z}(1-\zeta-\gamma)^{n-z-y}}{y!z!(n-y-z)!}. \qquad (1)$$

Note that the probability of a correct response is $\xi = \zeta+\gamma$ and that the probability of x correct responses to n randomly chosen items is

$$\binom{n}{x}\ \xi^{x}(1-\xi)^{n-x}.$$

This is the binomial error model described by Keats and Lord(1962).

To describe the distribution of true scores, $g(\xi)$, over the population of examinees, Keats and Lord (1962), Duncan (1974) and Morrison (1978) use the natural conjugate prior of the binomial distribution. That is, they assume

$$g(\xi) = \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)}\ \xi^{\nu_1-1}\ (1-\xi)^{\nu_2-1},$$

the beta distribution with parameters ν_1 and ν_2 where Γ is the usual gamma function. For some purposes this model has given good results with real data (Lord, 1965; Duncan, 1974). As a first approximation we consider the natural conjugate prior of (1) which is the bivariate Dirichlet distribution given by

$$g(\zeta,\gamma) = \frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1)\Gamma(\nu_2)\Gamma(\nu_3)}\,\zeta^{\nu_1-1}\gamma^{\nu_2-1}\,(1-\zeta-\gamma)^{\nu_3-1}$$

where $\nu_i>0, i=1, 2, 3$ and $\zeta+\gamma<1$. The marginal distributions of ζ and γ are beta distributions given by

$$g_1(\zeta) = \frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1)\Gamma(\nu_2+\nu_3)}\,\zeta^{\nu_1-1}\,(1-\zeta)^{\nu_2+\nu_3-1}$$

$$g_2(\gamma) = \frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_2)\Gamma(\nu_1+\nu_3)}\,\gamma^{\nu_2-1}(1-\gamma)^{\nu_1+\nu_3-1}.$$

For some situations (e.g., when the number of alternatives per item is large) it might be reasonable to require β, the probability of guessing, to be less than 1/2 which implies that γ is less than 1/2 as well. Thus, the model might be appropriate when the mean of γ is less than 1/2 and its variance is not too large so that the probability of γ greater than 1/2 is essentially equal to zero. Another and perhaps more satisfactory approach is to require $0<\gamma<1/2$ in a more direct fashion. Before considering how this might be done, it is useful to examine the present model in more detail.

A number of important properties about the distribution of y and z are immediately available from results reported by Ishii and Hayakawa (1960). In particular, the joint pdf of y and z is

$$f(y,z) = \frac{B(\nu_1+y,\nu_2+z,\ n+\nu_3-y-z)}{(n+2)(n+1)B(\nu_1,\nu_2,\nu_3)B(1+y,\ 1+z,\ n+1-y-z)}$$

where

$$B(\eta_1,\eta_2,\eta_3) = \frac{\Gamma(\eta_1)\Gamma(\eta_2)\Gamma(\eta_3)}{\Gamma(\eta_1+\eta_2+\eta_3)}$$

Also, the pdf of $x = y + z$ is

$$f(x) = \frac{B(\nu_1+\nu_2+x,\ n+\nu_3-x)}{(n+1)B(\nu_1+\nu_2,\nu_3)B(1+x,\ n+1-x)} \tag{2}$$

where

$$B(n_1,n_2) = \frac{\Gamma(n_1)\ \Gamma(n_2)}{\Gamma(n_1+n_2)}$$

is the usual beta function.

Another important distribution is the joint pdf of x and ζ. This can be derived as follows: From standard results on multinomial distributions we have that

$$f(x|\zeta,\gamma) = \binom{n}{x}(\zeta+\gamma)^x\ (1-\zeta-\gamma)^{n-x}.$$

Therefore, the joint pdf of x, ζ, and γ is

$$f(x,\zeta,\gamma) = \binom{n}{x}\ B^{-1}(\nu_1,\nu_2,\nu_3)\zeta^{\nu_1-1}\gamma^{\nu_2-1}(\zeta+\gamma)^x(1-\zeta-\gamma)^{n-x+\nu_3-1}.$$

Since

$$(\zeta+\gamma)^x = \sum_{w=0}^{x} \binom{x}{w}\gamma^w\zeta^{x-w}$$

and since for any non-negative integer w

$$\int \zeta^{x-w+\nu_1-1}\gamma^{w+\nu_2-1}(1-\zeta-\gamma)^{n-x+\nu_3-1}d\gamma$$

$$= B(w+\nu_2,n-x+\nu_3)\zeta^{x-w+\nu_1-1}\ (1-\zeta)^{n-x+w+\nu_2+\nu_3-1}$$

we see that the joint pdf of x and ζ is

$$\binom{n}{x}B^{-1}(\nu_1,\nu_2,\nu_3)\sum_{w=0}^{x}\binom{x}{w}\ B(w+\nu_2,n-x+\nu_3)\ \zeta^{x-w+\nu_1-1}\ (1-\zeta)^{n-x+w+\nu_2+\nu_3-1}. \tag{3}$$

Estimating ν_1, ν_2, and ν_3

We describe a method of estimating ν_1, ν_2 and ν_3 after which the motivation for using this estimation technique is discussed.

Consider N randomly sampled examinees and for the ith examinee let $\hat{\zeta}_i$ and $\hat{\beta}_i$ represent the maximum likelihood estimates of ζ and β. Note that $\hat{\gamma}=(1-\hat{\zeta})\hat{\beta}$ is a maximum likelihood estimate of γ. Let

$$\bar{\zeta}_j = N^{-1} \sum_i \hat{\zeta}_i^j \qquad (j=1,2)$$

$$\bar{\gamma}_j = N^{-1} \sum_i \hat{\gamma}_i^j \qquad (j=1,2)$$

$$\hat{\sigma}_\zeta^2 = \bar{\zeta}_2 - \bar{\zeta}_1^2$$

and

$$\hat{\sigma}_\gamma^2 = \bar{\gamma}_2 - \bar{\gamma}_1^2.$$

We estimate ν_1 and ν_2 with

$$\hat{\nu}_1 = \bar{\zeta}_1^2 (1-\bar{\zeta}_1)\hat{\sigma}_\zeta^{-2} - \bar{\zeta}_1$$

and

$$\hat{\nu}_2 = \bar{\gamma}_1^2 (1-\gamma_1)\hat{\sigma}_\gamma^{-2} - \bar{\gamma}_1$$

respectively. To estimate $\nu_2+\nu_3$ we use

$$\bar{\zeta}_1(1-\bar{\zeta}_1)^2\sigma_\zeta^{-2} + \bar{\zeta}_1 - 1$$

which in turn yields an estimate of ν_3.

The motivation for these estimates of the parameters of the Dirichlet prior stem from considering the asymptotic case. By the asymptotic case

we mean situations where both the number of item pairs used to estimate ζ and γ as well as the number of examinees tested are large.

The estimate of ζ for a given examinee is consistent and so as the number of item pairs gets large,

$$E(\hat{\zeta}^k|\zeta) \to \zeta^k \qquad (k=1,2).$$

Taking the expected value of both sides of this limit with respect to the distribution of ζ, we have from a result given by Rao (1973, p. 124) that

$$E(\hat{\zeta}^k) \to E(\zeta^k) \qquad (k=1,2)$$

If $E(\hat{\zeta}^k) = E(\zeta^k)$ which, as just indicated, is approximately true when the number of equivalent item pairs is sufficiently large, it follows from the Kolmogorov strong law of large numbers that

$$\bar{\zeta}_k \to E(\zeta^k)$$

In a similar manner $\bar{\gamma}_1$ and $\bar{\gamma}_2$ can be justified as an estimate of the moments of the distribution of γ. From known results about the beta-binomial distribution and since the marginal distribution of ζ and γ are beta distributions, it follows that expressions (2), (3) and (4) may be used to estimate ν_1, ν_2 and ν_3.

It is known that a large number of examinees might be required to obtain reasonably accurate estimates of the parameters of the beta-binomial model. Thus, a large number of examinees would seem to be prudent if accurate estimates of ζ and γ are desired.

Judging the Adequacy of the Model

At a minimum we want the model to give a good fit to $f(x)$, the marginal distribution of observed scores. As previously indicated, if β is set arbitrarily equal to zero, we get the beta-binomial error model. Thus, the model described here can be made to give a good fit to all 16 distributions reported by Lord (1965).

To further ascertain the adequacy of the model, we consider some of its implications. Consider first the relationship between ζ and β. Results reported by Connor and Mosimann (1969, p. 200) together with their Theorem 2 imply that ζ and $\beta = \frac{\gamma}{1-\zeta}$ are independent when a bivariate Dirichlet distribution for ζ and γ is assumed. Darroch and Ratcliff (1971) demonstrate that the converse is also true if, in addition, γ and $\zeta(1-\gamma)^{-1}$ are independent.

Under certain circumstances it might be reasonable to assume that ζ and β are uncorrelated and possibly independent (Swineford, 1941) but as pointed out by Frary (1969), this assumption might not be reasonable for many situations. However, if β is arbitrarily set equal to a constant (e.g., zero, or if guessing is assumed to be at random) this again implies that ζ and β are independent. Thus, insofar as interest centers on the proportion of skills an examinee has acquired, the present model is more "realistic" in that it allows β to vary over the population of examinees.

To allow for the possibility of ζ and β being correlated, one might replace the Dirichlet distribution of ζ and γ with the generalized Dirichlet distribution described by Connor and Mosimann (1969). This distribution arises if we assume γ and $\zeta(1-\gamma)^{-1}$ are independent (as is the case when the standard Dirichlet distribution is assumed) and that both have a univariate beta distribution. If, however, ζ is assumed to

have a beta distribution, the generalized Dirichlet distribution reduces to the standard Dirichlet distribution implying that ζ and β are again independent (James, 1972). In a more recent paper, James (1975) describes another generalization of the Dirichlet distribution which would allow ζ and β to be dependent but there are technical problems that preclude this distribution being used in practice.

Restricting the Range of β

As previously noted, there may be situations where it is reasonable to require $\beta<\frac{1}{2}$ which implies that γ is in this range as well. For the previously described distributions involving γ, it is possible to ensure that $0<\gamma<\frac{1}{2}$ by mapping γ into $\gamma/2$ and applying the standard change of variable technique. For example, $g_2(\gamma)$, the marginal distribution of γ, would now be

$$g_2(\gamma) = \frac{2\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1)\Gamma(\nu_1+\nu_3)} (2\gamma)^{\nu_2-1} (1-2\gamma)^{\nu_1+\nu_3-1}.$$

In otherwords, we are assuming that γ has a three parameter beta distribution (see, e.g., Johnson and Kotz, 1970) where the third parameter is arbitrarily set equal to 1/2 to ensure $0 < \gamma < \frac{1}{2}$. Note, however, that the marginal distribution of x, as well as the joint distribution of x and ζ, remains the same. Also, ζ and β are still independent.

Applications

We discuss two possible applications of the model after which a numerical illustration is given.

A New Formula Score

As previously indicated, several formula scores have been proposed in the literature which assume that guessing is at random when an examinee does not know the correct response. One advantage of the model described here is that it yields a new formula score which allows the probability of guessing to vary over the population of examinees. In particular, it is the conditional expectation of ζ given x which can derived as follows:

$$E(\zeta|x) = \frac{1}{f(x)} \int \zeta f(x, \zeta) d\zeta$$

$$= \frac{1}{f(x)} \binom{n}{x} B^{-1}(\nu_1, \nu_2, \nu_3) \sum_{w=0}^{x} \binom{x}{w} B(w+\nu_2, n-x+\nu_3) \int \zeta^{x-w+\nu_1} (1-\zeta)^{n-x+w+\nu_2+\nu_3-1}$$

$$= \frac{1}{f(x)} \binom{n}{x} B^{-1}(\nu_1, \nu_2, \nu_3) \sum_{w=0}^{x} \binom{x}{w} B(w+\nu_2, n-x+\nu_3) B(x-w+\nu_1+1, n-x+w+\nu_2+\nu_3). \quad (4.)$$

Application to Domain-Referenced Testing

Consider a specified domain of skills and suppose we want to know whether the proportion of skills an examinee knows is above or below a known constant, say $\zeta_0 = .7$. This question arises, for example, when using a domain-referenced test. Typically, however, the problem is stated in terms of ξ, the proportion correct true score of the examinee.

(See Hambleton et al., 1978, for a recent review of this area of measurement.) The present formulation of this problem would seem to be more interesting since we would expect the probability of guessing to be higher on certain tests than others. If two alternatives are used on each multiple-choice test item, rather than say 5, we would expect the observed scores on the first test to be higher than those on the second, particularly when a high proportion of examinees have not acquired the skills being tested. Thus, we might be able to make a test or a population of examinees "look good" by altering the number of alternatives that are used. By artificially changing the distribution of ξ, we might be able to ensure that true scores do not center around the criterion score ξ_0 which in turn might lower the probability of a false-positive or false-negative decision when comparing ξ to ξ_0. By formulating the problem in terms of ζ rather than ξ, we are attempting to eliminate this problem. One might also try to adjust ξ_0 according to the number of alternatives on each item but it is not immediately clear how this should be done unless perhaps we assume examinees guess at random.

For the sake of illustration we suppose that an extensive preliminary analysis of the item domain yields $\nu_1=10$, $\nu_2=3$, and $\nu_3=2$. The examiner is considering the use of a five-item test with a passing score of 4. That is, if $x \geq 4$, the decision $\zeta \geq \zeta_0$ is made. What is the probability of a false-positive decision, i.e., the probability of observing $x \geq 4$ when in fact $\zeta < \zeta_0 = .7$? This probability is given by

$$\sum_{x=4}^{5} \int_0^{.7} f(x,\zeta)d\zeta.$$

The evaluation of this quantity might appear to be formidable but it can be done fairly easily (once ν_1, ν_2, and ν_3 are estimated) using subroutine BDTR in IBM's Scientific Subroutine Package (1971). For the example described above the estimate of a false-positive decision is .4623. This would seem to be intolerably high for most applications. One solution is to increase the passing score to 5. In this case the probability is .264. The probability of a false-negative decision will increase, however, the value of which can be determined in a similar manner. Another method of lowering the probability of a false-positive decision is to increase the number of items on the test.

The Case $\beta=0$, $\alpha>0$

When $\beta=0$ and $\alpha>0$, an examinee belongs to one of three disjoint sets in terms of a particular achievement test item:

1. The examinee knows the answer and gives an incorrect response.
2. The examinee does not know and is incorrect.
3. The examinee knows and is correct.

The probabilities associated with these three events are $\zeta\alpha$, $(1-\zeta)$, and $\zeta(1-\alpha)$, respectively. To obtain a model similar to the one described above we set $\gamma_1=\zeta\alpha, \gamma_2=1-\zeta$ and say that γ_1 and γ_2 arise from a bivariate Dirichlet distribution. The derivation of the results relating γ_1 and γ_2 to the observed scores is essentially the same as the case $\beta>0$, $\alpha=0$ and so it is not discussed.

An Illustration

For the sake of illustration, test data was generated for 1000 examinees responding to 100 pairs of equivalent items. This was accomplished as follows: For each examinee a value for ζ and β was generated by setting $\nu_1=8$, $\nu_2=3$, and $\nu_3=2$. The resulting beta distribution for both α and β was then approximated by a step function having 100 steps. This step function was then used to generate ζ and β in conjunction with the random number generator RANDU (IBM, 1971). Once α and β were determined, RANDU was again used to generate the observed responses of the examinee to the 100 pairs of equivalent items.

The resulting estimates of ν_1, ν_2 and ν_3 were $\hat{\nu}_1=6.6$, $\hat{\nu}_2=1.5$ and $\hat{\nu}_3=2.7$. Although hardly conclusive, this suggests that reasonably accurate estimates of ν_1, ν_2 and ν_3 are possible when a large (but realistic) number of items and examinees are available. The main point of this illustration is that once ν_1, ν_2 and ν_3 are estimated, it is possible to modify the test to meet certain practical demands of a given situation. Suppose, for example, we want to limit a test to 20 items, no two of which are equivalent. Since the values of ν_1, ν_2 and ν_3 are the same for different values of n, we may correct for guessing by applying equation (4). The resulting estimates of ζ for each of the 21 possible observed scores are summarized in Table 2.

REFERENCES

Abu - Sayf, F. K. A new formula score Educational and Psychological Measurement, 1977, 37, 853-862.

Anderson, T. W. Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.) Probability and Statistics. New York: John Wiley and Sons, 1959.

Barnett, V. D. Evaluation of the maximum - likelihood estimator where the likelihood equation has multiple roots. Biometrika, 1966, 53, 151-165.

Brownless, V. T. and Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Chernoff, H. The scoring of multiple choice questionnaires. Annals of Mathematical Statistics, 1962, 33, 375-393.

Connor, R. J. and Mosimann, J. E. Concepts of independence for proportions with a generalization of the Dirichlet distribution. Journal of the American Statistical Association, 1969, 64, 194-206.

Darroch, J. N. and Ratcliff, D. A characterization of the Dirichlet distribution. Journal of the American Statistical Association, 1971, 66, 641-643.

Dayton, C. M. and Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Diamond, J. and Evans, W. The correction for guessing. Review of Educational Research, 1973, 43, 181-191.

Duncan, George T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

Frary, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. Educational and Psychological Measurement, 1969, 29, 359-365.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hamilton, C. H. Bias and error in multiple-choice tests. Psychometrika, 1950, 15, 151-168.

IBM Application Program, System/360. Scientific subroutines package (360-CM-03X) Version III, programmer's manual. White Plains, New York: IBM Corporation Technical Publications Department, 1971.

Ishii, G. and Hayakawa, R. On the compound binomial distribution. Annals of the Institute of Statistical Mathematics, 1960, 12, 69-80.

James, I. R. Products of independent beta variables with application to Connor and Mosimann's generalized Dircihlet distribution. Journal of the American Statistical Association, 1972, 67, 910-912.

James, I. R. Multivariate distributions which have beta conditional distributions. Journal of the American Statistical Association, 1975, 70, 681-684.

Johnson, N. and Kotz, S. Continuous Univariate Distributions - 2, New York: John Wiley, 1970.

Kale, B. K. On the solution of the likelihood equation by iteration processes. Biometrika, 1961, 48, 452-456.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Knapp, T. R. The reliability of a dichotomous test-item: A "correlationless" approach. Journal of Educational Measurement, 1977, 14, 237-252.

Lazarsfeld, P. F. and Henry, N. W. Latent structure analysis New York: Houghton Mifflin, 1968.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Mass: Addison - Wesley, 1968.

Lord, F. M. Formula scoring and number-right scoring. Journal of Educational Measurement, 1975, 12, 7-11.

Lyerly, S. B. A note on correcting for chance success in objective tests. Psychometrika, 1951, 16, 21-30.

Macready, G. B. and Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marks, E. and Noll, G. A. "Procedures and criteria for evaluating reading and listening comprehension tests." Educational and Psychological Measurement, 1967, 27, 335-348.

McNemar, Q. Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 1947, 12, 153-157.

Morrison, D. G. A probability model for forced binary choices. The American Statistician, 1978, 32, 23-25.

Proctor, C. H. A probabilistic formulation and statistical analysis of Guttman scales. Psychometrika, 35, 73-78.

Rao, C. R. Linear statistical inference and its applications. New York: Wiley, 1973.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 32, 438-444.

White, R. T. and Clark, R. M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. Pearlman and R. Wilcox Achievement Test Items: Methods of Study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977(a).

Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. Pearlman, and R. Wilcox, Achievement Test Items: Methods of Study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977(b).

Wilcox, R. R. An alternative interpretation of three stability models. Educational and Psychological Measurement, in press.

Table 1

The Probabilities Corresponding to the Four Possible Outcomes of an Examinee Responding to a Randomly Chosen Item.

	Correct	Incorrect
Knows	$\zeta(1-\alpha)$	$\zeta\alpha$
Does Not Know	$(1-\zeta)\beta$	$(1-\zeta)(1-\beta)$

Table 2

Regression Estimate of ζ for an observed score of x, n=20.

x:	0	1	2	3	4	5	6	7	8	9	10
$E(\zeta\|x)$:	.21	.24	.27	.29	.32	.35	.37	.40	.43	.45	.48

x:	11	12	13	14	15	16	17	18	19	20
$E(\zeta\|x)$:	.51	.53	.56	.58	.61	.64	.66	.69	.72	.74